DOCUMENT RESUME

ED 118 607 TM 005 094

AUTHOR Livingston, Samuel A.
TITLE A Utility-Based Approach to the Evaluation of Pass/Fail Testing Decision Procedures. COPA-75-01.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-COPA-75-01
PUB DATE Jul 75
NOTE 12p.

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage
DESCRIPTORS *Cutting Scores; *Decision Making; Mathematical Models; Measurement Techniques; Statistical Analysis; *Testing; Test Validity
IDENTIFIERS Pass Fail Testing; Utility Ratio

ABSTRACT

A measure of the usefulness of a pass/fail testing decision procedure is the ratio of the utility of the given procedure to the utility of a procedure based on knowledge of scores on a criterion measure. It is computed from scores for a representative sample of persons tested. Utility functions may be specified by the test user or set by convention to be linear with unit slope. The utility ratio can be used for comparing tests or for selecting test items. (Author)

COPA-75-01

A Utility-Based Approach to the Evaluation of Pass/Fail Testing Decision Procedures

Samuel A. Livingston

This Bulletin is a draft for interoffice circulation. Corrections and suggestions for revision are solicited. The Bulletin should not be cited as a reference without the specific permission of the author. It is automatically superseded upon formal publication of the material.

Center for Occupational and Professional Assessment
Educational Testing Service
Princeton, New Jersey

July 1975

Consider the following situation: A decision-maker intends to use a test to make a decision about each of many persons. For each person the decision-maker must take one of two possible actions, which we will call "action A" and "action R". The decision-maker considers action A more appropriate for persons who score high on the test and action R more appropriate for low scorers; the letters A and R might stand for "accelerated program" and "regular program", or "award credit" and "refuse credit", or simply "accept" and "reject". The decision-maker will choose one specific point on the test-score continuum as the cutoff point; all persons with test scores at or above this point will receive action A, while all those with test scores below the cutoff will receive action R. We will use the symbol x_o to refer to this cutoff score. We will restrict our attention to the situation in which there are no constraints on the numbers of persons assigned to actions A and R.

Now let us suppose that the decision-maker would like to validate this decision procedure against a criterion measure (either concurrently or predictively), by administering the criterion measure to a representative sample of persons who have taken the test. The higher a person's score on the criterion measure, the better the result of action A for that person and the worse the result of action R. At some point on the criterion-measure scale, the decision-maker would be undecided between actions A and R. We will use the symbol y_o to refer to this indifference point.[1]

---

[1] The choice of y_o is logically prior to the choice of x_o. Procedures for optimizing the choice of x_o for a given value of y_o are discussed by Davis, Hickman, and Novick (1973).

We can express our decision-maker's feelings mathematically in the form of two utility functions. Let y_i represent the score of person i on the criterion measure. Then let

$u_a(y_i)$ = utility of action A for person i,

$u_r(y_i)$ = utility of action R for person i

where $u_a(y_o) = u_r(y_o) = 0$. That is, the zero on the utility scale is defined to be the value of either action at the indifference point. We assume that u_a is an increasing function and u_r a decreasing function, to reflect the greater importance of correct decisions about persons whose criterion performance is farther from the indifference point.[2] Figure 1 presents an example of a possible pair of utility functions. (Note that the criterion measure is plotted along the horizontal axis.)

Our decision-maker intends to use a decision procedure that can be expressed mathematically as follows: Let x_i be the test score of person i and let x_o be the minimum passing score. Then we take action A for person i if $x_i \geq x_o$ and action R if $x_i < x_o$. The utility of this decision procedure is the sume of the utilities of all the individual decisions:

$$U(x_o) = \sum_{x_i \geq x_o} u_a(y_i) + \sum_{x_i < x_o} u_r(y_i)$$

As a standard for comparison we have the utility of the ideal decision procedure based on knowledge of each person's performance on the criterion measure:

$$U(y_o) = \sum_{y_i \geq y_o} u_a(y_i) + \sum_{y_i < y_o} u_r(y_i)$$

---

[2] This feature of the situation distinguishes it from the "threshold utility" situation examined by Hambleton and Novick (1973) and by Petersen (1974).

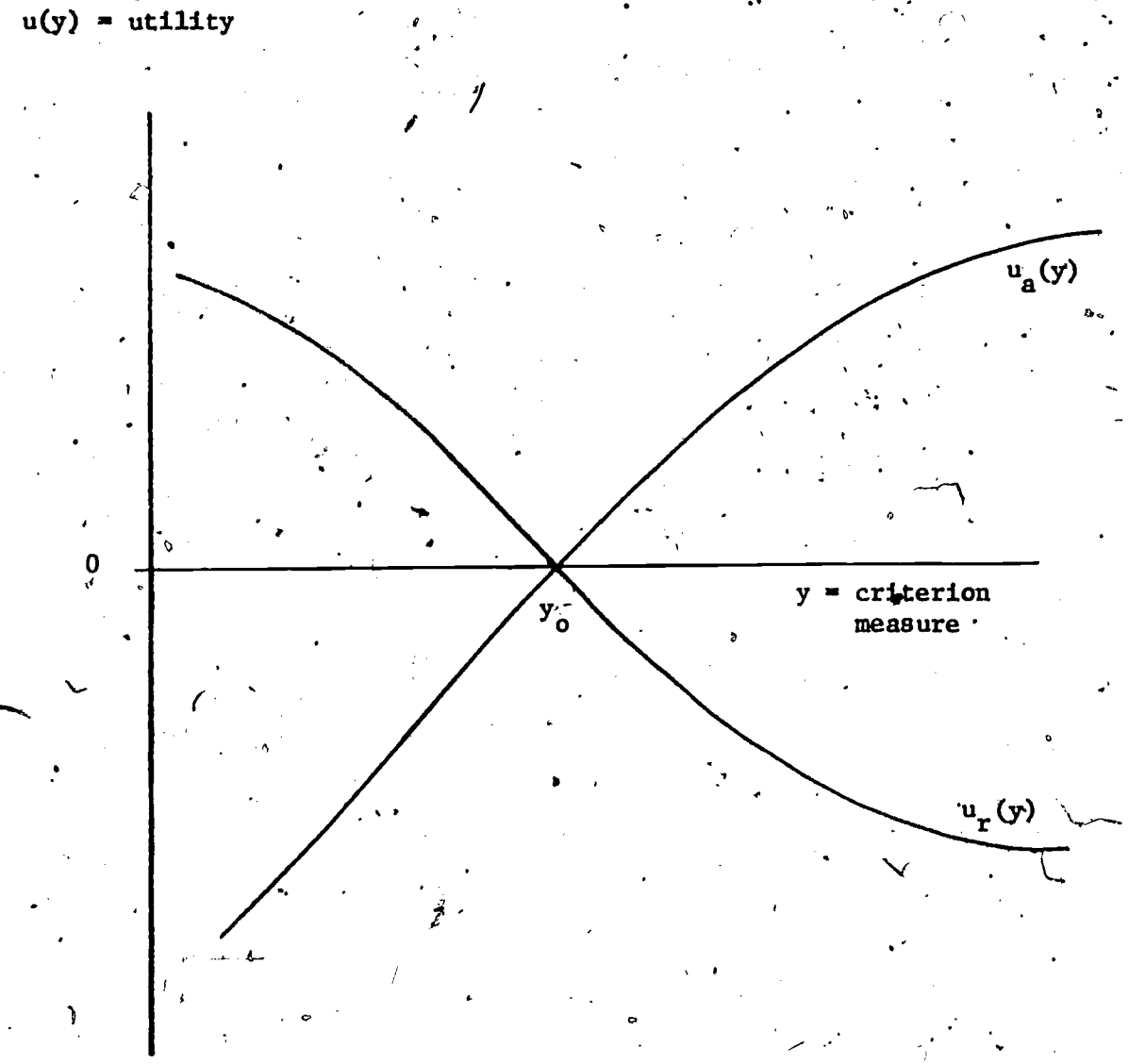

Figure 1. Illustration of one possible choice of utility functions.

Then the validity of the decision procedure based on x_o, as compared to the validity of the "ideal" decision procedure based on y_o, can be described by the ratio

$$U(x_o, y_o) = \frac{U(x_o)}{U(y_o)}$$

This utility ratio[3] is expressed as a function of the minimum passing score x_o and the indifference point y_o to emphasize its dependence on the choice of x_o and y_o.

Because the denominator of the utility ratio is the maximum utility over all possible sets of decisions -- the utility of a correct decision for every person -- the utility ratio reaches its maximum at 1. The minimum value for the utility ratio is not necessarily -1 unless $u_r(y) = -u_a(y)$ for all values of y. The utility ratio equals zero when the harm from the bad decisions exactly balances the benefit from the good decisions. A negative utility ratio indicates that the decision procedure could have been improved by taking action A for the low scorers and action R for the high scorers. (This situation would be expected if the test were accidentally reverse-scored.)

One type of utility function that is of particular interest because of its simplicity and intuitive appeal is that represented by straight lines. Let b_a be the benefit of accepting a person one unit above y_o on the criterion measure and let c_r be the cost of rejecting that person. Similarly, let b_r be the benefit of rejecting and c_a be the cost of accepting a person one unit below

---

[3] This ratio does not correspond to the "utility ratio" defined for the threshold utility case by Petersen (1974). Petersen's utility ratio does not depend on observed data, but merely describes the utility functions.

y_o on the criterion measure. Then let

$$u_a(y) = b_a (y - y_o) \text{ if } y \geq y_o$$
$$c_a (y - y_o) \text{ if } y < y_o$$

$$u_r(y) = -b_r (y - y_o) \text{ if } y < y_o$$
$$-c_r (y - y_o) \text{ if } y \geq y_o$$

Utility functions of this form imply that the cost of a bad decision is proportional to the size of the error. Similarly, they imply that the benefit from a good decision is proportional to the size of the error that was avoided. The size of the error made or avoided is the absolute value of $(y_i - y_o)$.[4] These utility functions could be described as "semi-linear"; they become fully linear when $b_a = c_a$ and $b_r = c_r$.

Figure 2 illustrates a pair of utility functions of this form. Only the relative sizes of b_a, c_a, b_r, and c_r affect the value of the utility ratio, as can be seen by multiplying all four coefficients by any constant k. This multiplication would have the effect of multiplying both utility functions by k. Therefore the numerator and denominator of the utility ratio would both be multiplied by k, leaving its value unchanged.

What is the expected utility of a decision procedure in which actions A and R are assigned purely at random? Is it necessarily zero? Let p_a and p_r be the probabilities of assigning actions A and R, respectively. Then the expected utility of the decision procedure is

$$\sum_{\text{all } i} [p_a u_a(y_i) + p_r u_r(y_i)]$$

---

[4] The coefficients b_a, c_r, b_r, and c_a correspond to Petersen's (1974) utility values a, b, c, and d, respectively, except that in Petersen's approach they are not multiplied by the size of the error.

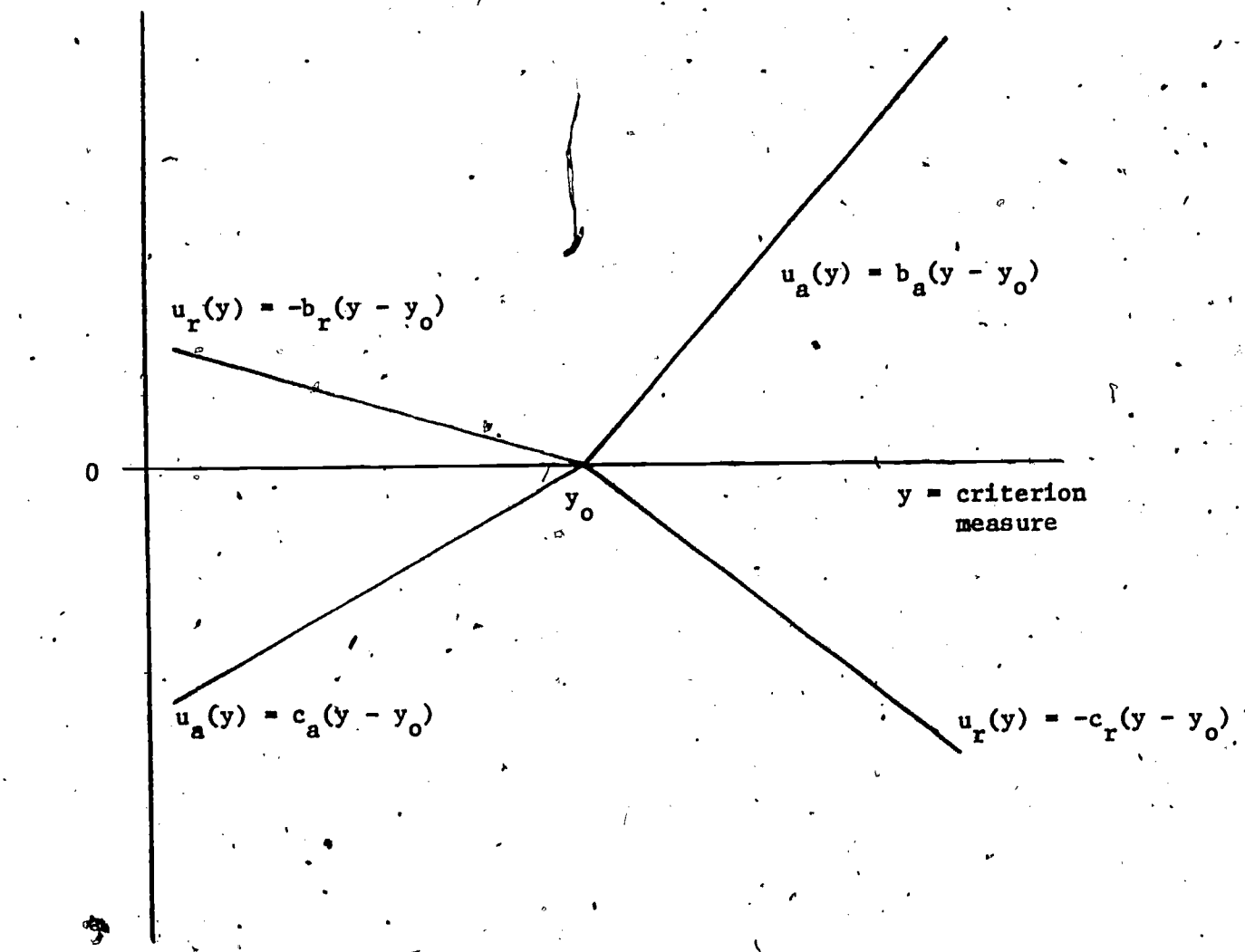

Figure 2. Illustration of "semi-linear" utility functions.

This expression will equal zero when the utility functions are such that $u_a / (-u_r) = p_r / p_a$ for all values of y. If the utility functions are fully linear,

$$u_a(y) = b_a(y - y_o)$$

$$u_r(y) = -b_r(y - y_o) ,$$

then the expected utility of the random decision procedure will be positive when

$$\sum_{\text{all } i} [p_a \; b_a (y - y_o) - p_r \; b_r (y - y_o)] > 0;$$

that is, when $(p_a b_a - p_r b_r)$ has the same sign as $(\bar{y} - y_o)$. Therefore, if the average score on the criterion measure were far enough above the indifference point and the benefit or harm from action A sufficiently greater than that from action R, the decision-maker would do reasonably well by taking action A for all persons. (This example shows the importance of the requirement that the validation sample be representative of the group of persons about whom decisions are to be made.)

Why should a test user such as the decision-maker described at the beginning of this paper use the utility ratio for evaluating his test-based decision procedure on the basis of a criterion measure? Wouldn't one of the more familiar correlation-like statistics serve his purpose just as well? No, because none of the more familiar statistics uses all that information and only that information that the decision-maker actually uses in making his decisions and evaluating their results. The utility ratio treats the test score as a dichotomous variable

because the test score is being used as a dichotomous variable. At the same time, it does not impose an unnecessary dichotomy on the criterion measure, as do the phi-coefficient and the per-cent-agreement statistic. While it treats the criterion measure as continuous, it takes into account the indifference point that forms a natural zero for the criterion measure and thus makes it a meaningful ratio scale. Finally, it allows the decision-maker to adopt whatever utility functions best reflect his values.

Traditionalists may object that a utility-based approach to test validation allows the decision-maker too much freedom to influence the value of the resulting coefficient. This objection can be overcome by establishing a convention of computing utility ratios on the basis of fully linear utility functions with equal slopes: $u_a(y) = y - y_o$; $u_r(y) = -(y - y_o)$. An administrator or researcher who proposes a different set of utility functions in a particular situation would then be obligated to show why the utility functions he advocates are more appropriate than those established by convention.[5]

The most obvious use of the utility ratio is for comparing two or more tests. However, it also offers a practical alternative to the use of traditional discrimination indices for selecting test items for a test intended to discriminate at a particular level of ability (either on an external criterion variable or on the test itself). It also allows the test constructor to specify the relative importance of identifying qualified versus unqualified examinees. Let $x_i = 1$ if the examinee answers the item correctly and $x_i = 0$ if he does not. Then if an external criterion variable is used as the basis for item selection, the utility ratio

---

[5] Notice that whenever we use a traditional product-moment correlation to validate a test, we implicitly accept the convention that the utility of test score x for a person with criterion value y is given by the product $(x - \bar{x})(y - \bar{y})$.

for a given item would be

$$\frac{\sum_{x_i = 1} u_a(y_i) + \sum_{x_i = 0} u_r(y_i)}{\sum_{y_i \geq y_o} u_a(y_i) + \sum_{y_i < y_o} u_r(y_i)}$$

If scores on the test itself were used as the basis for item selection, the test constructor would have to specify utility functions in terms of test scores. In this case the y's in the above formula would refer to scores on the full test; y_o would represent the score level at which maximum discrimination is desired.

Acknowledgments

I wish to thank Thomas F. Donlon, Frederick R. Kling, Michael M. Ravitch, and Cheryl Wild Reed for their many helpful comments on earlier drafts of this paper.

REFERENCES

Davis, C. E., Hickman, Jr., and Novick, M. R. A primer on decision analysis for individually prescribed instruction. Technical Bulletin No. 17. Iowa City, Iowa: American College Testing Program, 1973.

Hambleton, R. L., and Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Petersen, N. S. An expected utility model for "optimal" selection. Technical Bulletin No. 24. Iowa City, Iowa: American College Testing Program, 1974.